March 24, 2023

C964: Computer Science Capstone

Ryan Blumenhorst

Task 2 parts A, B, C, and D

[Part A: Project Proposal for Business Executives 4](#_Toc1515453419)

[Letter of Transmittal 4](#_Toc1044087116)

[Project Recommendation 6](#_Toc1721750458)

[Problem Summary 6](#_Toc205640471)

[Application Benefits 6](#_Toc739883765)

[Application Description 6](#_Toc1967641075)

[Data Description 7](#_Toc1580263866)

[Objectives and Hypothesis 7](#_Toc1084915425)

[Methodology 8](#_Toc1455093807)

[Funding Requirements 8](#_Toc1716359054)

[Data Precautions 8](#_Toc821607173)

[Developer’s Expertise 8](#_Toc742297278)

[Part B: Project Proposal 9](#_Toc1516064891)

[Problem Statement 9](#_Toc602847551)

[Customer Summary 9](#_Toc411678477)

[Existing System Analysis 9](#_Toc1668713332)

[Data 9](#_Toc1911399030)

[Project Methodology 10](#_Toc1307313575)

[Project Outcomes 10](#_Toc1967172832)

[Implementation Plan 11](#_Toc306674583)

[Evaluation Plan 11](#_Toc849271423)

[Resources and Costs 12](#_Toc1695791095)

[Timeline and Milestones 12](#_Toc787992355)

[Part C: Application 14](#_Toc1767968421)

[Part D: Post-implementation Report 15](#_Toc22579049)

[A Business (or Organization) Vision 15](#_Toc1496854501)

[Datasets 16](#_Toc645781746)

[Data Product Code 17](#_Toc157285634)

[Objective (or Hypothesis) Verification 18](#_Toc1315546441)

[Effective Visualization and Reporting 18](#_Toc656892821)

[Accuracy Analysis 20](#_Toc1956352798)

[Application Testing 21](#_Toc871280264)

[Application Files 21](#_Toc1639588139)

[User Guide 22](#_Toc1125023771)

[Summation of Learning Experience 22](#_Toc1702287338)

# Part A: Project Proposal for Business Executives

## Letter of Transmittal

Date: March 22, 2023

To: R. Blume, Owner & Operator, BlumeHospitals

From: Ryan Blumenhorst, CEO, BlumeTech

Subject: New Solution for Early Parkinson’s Disease Detection

Dear Mr. Blume,

In the past few months, BlumeTech has branched into hospitals and clinics all around the United States. Our machines provide the best in accurate diagnosis for many different illnesses. However, one illness that we haven’t used our technology for is Parkinson’s Disease. Parkinson’s Disease, or PD, affects many people and can be hard to detect early. Early detection of PD is crucial to getting our patients the therapy that they need to keep their motor functions for as long as possible. Human detection of PD, expecially early detection, can sometimes be innaccurate. Thus, our team here at BlumeTech has created a solution to help detect PD earlier than humans possibly could.

Our solution, the Parkinson’s Vocal Detection Machine, uses a newer algorithm of machine learning. This algorithm is called XGBoosting, or eXtreme Gradient Boosting. This algorithm is based on decision trees and is designed to be fast and easy to implement. Using this newer algorithm, we can guarentee higher accuracy than a human doctor could provide. But how exactly does this algorithm work? This algorithm will use the biomedical vocal data of patients with and without PD and will be able to use new patients vocal data to accurately detect PD in the patient.

But how much would this solution cost our company? Well with how this solution is made, we would only need to provide a machine to run the program and record the vocal data and time to train employees. To fully implement this solution, we are estimating around $3,000. This will provide our main location with this new and accurate solution.

You may still have some questions about us and our application. Our company firmly believes in using technology for the betterment of human health. We have provided many medical machines to different hospitals and clinics around the United States. Our company has provided many unique solutions to medical software and hope to expand into BlumeHospitals to help patients with PD get the therapy that they need as early as possible. We hope that you share the same vision.

We look forward to hearing from you soon!

All the best,



Ryan Blumenhorst, CEO BlumeTech

## Project Recommendation

### Problem Summary

With this AI-Based system, we will be able to detect the presence of Parkinson’s Disease within a human. It will be able to do so by using biomedical vocal readings caught by a machine and using that data to give an accurate detection of Parkinson’s Disease. The detection of PD can be difficult to catch early in the diseases lifespan. Human eyes often cannot see the signs of PD until it progresses further into the disease. Your hospital focuses on the best care for your patients. With this project, you will have access to our machine that will have a 94% accurate reading for PD detection. This solution will give you and your patients the early knowledge of PD so that your patients can get to the therapy that they need sooner.

### Application Benefits

The detection of Parkinson’s Disease can sometimes be difficult to catch early on. This is because traditional diagnostic approaches rely on the evaluation of movements that are sometimes too subtle for human eyes to see. This can lead to difficulty detecting Parkinson’s or even the misdiagnosis of Parkinson’s. With this machine learning project, we will be able to accurately detect these movements and help detect the presence of Parkinson’s Disease.

Your organization needs this machine learning project to reduce the number of human errors when detecting Parkinson’s Disease. With this machine learning program, you will be able to detect Parkinson’s Disease earlier than a human could and can quickly get your patients the therapy that they will need and provide better service.

### Application Description

With this new AI, your organization will be able to accurately detect whether a person has Parkinson’s Disease. This AI will be able to detect Parkinson’s Disease in its early stages, which is critical for the patients. This will improve the accuracy of diagnosis and provide your patients with the quickest access to the physical therapy that they will need.

This AI solution will use a biomedical voice dataset from UCI to train the program to read the measurements that are collected from the patient. These measurements include average vocal fundamental frequency, maximum and minimum vocal fundamental frequency, measurements of the ratio of noise to tonal components in the voice and many other readings. These readings will be used to train the machine learning algorithm using XGBoost (eXtreme Gradient Boost), an algorithm based on decision trees that has been created to be fast and reliable. 80% of the dataset will be used to train the program, and the last 20% will be used to test the algorithm. After the program has been trained, the solution will have an accuracy rating of 94%. At this point, the algorithm will be able to take the input of new data and accurately detect Parkinson’s Disease. This solution can be implemented quickly into our systems by providing a small computer that will take the readings from vocal data and input it into the algorithm. This machine learning project will become more reliable than human readings for Parkinson’s Disease.

### Data Description

For this AI, the dataset that will be used to train the AI will be the UCI ML Parkinson’s dataset (available from the website https://archive.ics.uci.edu/ml/datasets/parkinsons). This dataset contains 195 records with over 20 measurement readings. The other sources of data will come from the patients testing and readings using specialized instruments.

Data will be collected from patient testing and biomedical voice measurement readings. This data will then be entered into the AI and the machine will try to detect Parkinson’s Disease. An advantage of this data collection method is that if Parkinson’s is detected, the patient will have an instant reading and will then have access to the physical therapy that they will need. A limitation of this data collection, however, is that errors in the measurements can occur, such as if the sensor devices aren’t in the correct locations.

The data will be prepared in CSV format. The data will be a collection of different measurements obtained from the patients. If the data is missing, that data will not be able to be used in this algorithm. An error will be in place to let the user know that not all of the measurements have been plugged into the system and that the user must re-measure the readings. Any outliers would have to be re-measured to ensure an accurate reading.

### Objectives and Hypothesis

Objectives:

* Implement a XGBoost machine learning algorithm with classification accuracy of above 80%.
* Install and ready to use within two weeks
* Provide a more precise model of detecting Parkinson’s than human clinical detection (80% classification accuracy)
* Model will be accurate and cost efficient at detecting Parkinson’s Disease

My hypothesis is that using this machine learning algorithm, we will have an accurate and reliable program to detect Parkinson’s Disease earlier than a human could. The testing that will be done to prove this will be using real patients readings, and using their readings, accurately determine if the patient has Parkinson’s Disease. My prediction accuracy will be greater than 80%.

### Methodology

This project will be using the Agile framework. Continuous testing and feature integration are important for the success of this application. All feedback from the testing will be collected and considered to be integrated into the second phase deployments. Using the Agile framework, we have adopted the use of sprints. We will break the application into key points in order to achieve a successful application in a timely manner. The first sprint will focus on porting the application into an executable software. The second sprint will focus on integrating the application into the existing software programs. The final sprint will be the application testing and rollout.

### Funding Requirements

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Human Resources | Programmer | $500 |
| Electronic Devices | Machine to run program | $1,000 |
| Work Hours to Implement | Time to implement the new device | $1,500 |
|  | **Total** | $3,000 |

### Data Precautions

The data that is being collected and used is sensitive patient data. This data should be handled carefully to abide by patient confidentiality. Only authorized personnel should have access to this data. This can be done by having two-factor authorization (Providing passcode and badge/key) before this machine AI would be able to use. As a precaution, this machine will not save the input data onto its system. The readings will be moved to another secure database that keeps patient information, then when the machine is shut off, the input data is deleted. These forms of authorization are already in use by our organization as we handle other forms of sensitive patient data. Communication of this patient data is prohibited outside of doctor and patient. No sensitive patient data, such as the readings from this machine, should be discussed with others.

### Developer’s Expertise

I will be developing this solution. I have a Bachelor’s Degree in Computer Science from WGU and 4 years of professional software development. The software that I have been a part in making have all been machine learning projects. These machine learning projects have prepared me for this project.

# Part B: Project Proposal

## Problem Statement

Early detection of Parkinson’s Disease by humans can be innaccurate. This innaccuracy can lead to misdiagnosis or patients not having access to the therapy that they need as soon as possible. This AI-based solution will provide a more accurate detection of Parkinson’s Disease in patients.

## Customer Summary

The customers, in this case the patients, need to have the best care possible. This solution will be able to provide customers with an accurate reading of whether or not they need to go into Parkinson’s Therapy or the peace of mind that they don’t have Parkinson’s Disease.

## Existing System Analysis

Currently, you are using human detection of Parkinson’s Disease. The problems that can arise from this is that human detection of Parkinson’s Disease averages around 80%. However, the earlier the disease is, the harder it is to detect and the lower the human accuracy drops. Our solution will provide an accuracy of 94%, no matter how far along the patient is in their Parkinson’s. This solution will provide a more accurate reading and get the patients the therapy that they need earlier than a human might diagnose.

## Data

For this AI, the dataset that will be used to train the AI will be the UCI ML Parkinson’s dataset (available from the website https://archive.ics.uci.edu/ml/datasets/parkinsons). This dataset contains 195 records with over 20 measurement readings. The other sources of data will come from the patients testing and readings using specialized instruments.

Data will be collected from patient testing and biomedical voice measurement readings. This data will then be entered into the AI and the machine will try to detect Parkinson’s Disease. An advantage of this data collection method is that if Parkinson’s is detected, the patient will have an instant reading and will then have access to the physical therapy that they will need. A limitation of this data collection, however, is that errors in the measurements can occur, such as if the sensor devices aren’t in the correct locations.

The data will be prepared in CSV format. The data will be a collection of different measurements obtained from the patients. If the data is missing, that data will not be able to be used in this algorithm. An error will be in place to let the user know that not all of the measurements have been plugged into the system and that the user must re-measure the readings. Any outliers would have to be re-measured to ensure an accurate reading.

The data that is being collected and used is sensitive patient data. This data should be handled carefully to abide by patient confidentiality. Only authorized personnel should have access to this data. This can be done by having two-factor authorization (Providing passcode and badge/key) before this machine AI would be able to use. As a precaution, this machine will not save the input data onto its system. The readings will be moved to another secure database that keeps patient information, then when the machine is shut off, the input data is deleted. These forms of authorization are already in use by our organization as we handle other forms of sensitive patient data. Communication of this patient data is prohibited outside of doctor and patient. No sensitive patient data, such as the readings from this machine, should be discussed with others.

## Project Methodology

The development of this project will follow the SEMMA methodology.

* **Sample**: We will sample a dataset consisting of 195 records and only 39.7 KB in size. This will be large enough to get accurate training, while also being small in storage size.
* **Explore**: We will explore this data set and the over 20 measurements that are recorded to better understand what each of these measurements mean and how they can be used to detect Parkinson’s Disease.
* **Modify**: We will modify the data by creating new variables for these measurements and see what the AI can determine from the new dataset.
* **Model**: We will model the data efficiently so that we can quickly and reliably determine the desired outcome.
* **Assess**: We will assess how accurate the model is and if it is reliable and usable.

## Project Outcomes

Goals:

* Provide an accurate reading of whether or not a person has Parkinson’s Disease.
* Accurately and securely use new data
* Project will cost less than $5,000
* Project will be completed in under 2 months

Objectives:

* Implement a XGBoost machine learning algorithm with classification accuracy of above 80%.
* Install and ready to use within two weeks
* Provide a more precise model of detecting Parkinson’s than human clinical detection (80% classification accuracy)
* Model will be accurate and cost efficient at detecting Parkinson’s Disease

Deliverables:

* A machine learning algorithm to detect Parkinson’s Disease.
* Documentation on how the machine learning algorithm works. (User Manual)

## Implementation Plan

* Provide an outline of how the project will be implemented. This description might include the following:
  + General strategy.
  + Phases of the rollout.
  + Dependencies.
  + Details for testing and distribution.

The implementation plan focuses on two phases and the rollout of the prototype and final application. At the end of each phase, the application and all associated resources will be rolled out to the main facility. These resources comprise of a brief training video and two weeks of initial support. We will also include a set of training data to ensure that users can practice using that data set.

We will restrict the code repository to only members of the development team at the hospital. This will prevent any edits to the code base by other employees and ensure proper change control management and that no developer can push changes without approval.

Programming environment:

* Python 3.10 along with pandas, numpy, PySimpleGUI, matplotlib, sklearn, and xgboost libraries.

## Evaluation Plan

Performance will be measured by how accurately (Classification Accuracy) the AI can detect Parkinson’s Disease in patients. This can be done by using data from known patients with Parkinson’s Disease and patients without and determining how accurately the AI can detect who has Parkinson’s Disease. This algorithm states that its training classification accuracy is 94%. This was found by splitting the dataset used to train the program (195 records) into an 80-20 split. 80% of the data (156 records) was used to train the algorithm, and the remaining 20% of data (39 records) was used to test the algorithm. To be as accurate as possible is the goal, so we won’t want the algorithm to be less than 80% accurate when testing with a larger group (above 50 records). The algorithm accuracy will fluctuate depending on the size of the dataset, so as long as the algorithm can stay above 80% classification accuracy, it would be considered successful. We can test the accuracy of the algorithm by providing a larger dataset of known Parkinson’s patients and known non-Parkinson’s patients. The maximum allowed error in this model would be 20%. If the algorithm has a classification accuracy of less than 80%, the model will be underperforming human clinical detection (Rizzo, 2016).

Performance will also be measured on how quickly the algorithm can run and produce the desired results. Our goal is to be able to accurately detect Parkinson’s but also quickly detect it as well. This algorithm will be tested with larger data sets and should take less than 5 seconds to produce the desired results.

## Resources and Costs

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Human Resources | Programmer | $500 |
| Electronic Devices | Machine to run program | $1,000 |
| Work Hours to Implement | Time to implement the new device | $1,500 |
|  | **Total** | $3,000 |

## Timeline and Milestones

3-10-2023: The proposal is accepted.

3-20-2023: A technical proof of machine learning AI detecting Parkinson’s Disease

3-22-2023: AI submitted for review.

3-25-2023: Creation of the software in a deliverable state. Program ready to install in select locations.

3-30-2023: AI Delivered to select locations.

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **Start** | **End** | **Tasks** |
| 1 | 3-10-2023 | 3-20-2023 | Create a technical proof of a Machine Learning AI detecting Parkinson’s Disease |
| 2 | 3-20-2023 | 3-22-2023 | Submit AI for review |
| 3 | 3-22-2023 | 3-25-2023 | Create the software in a deliverable state and ready to install |
| 4 | 3-25-2023 | 3-30-2023 | Deliver and install AI in selected locations |

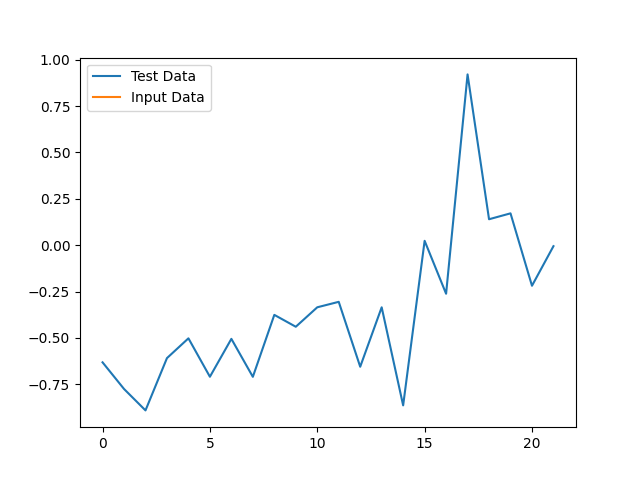
# Part C: Application

The application has been provided along with this documentation.

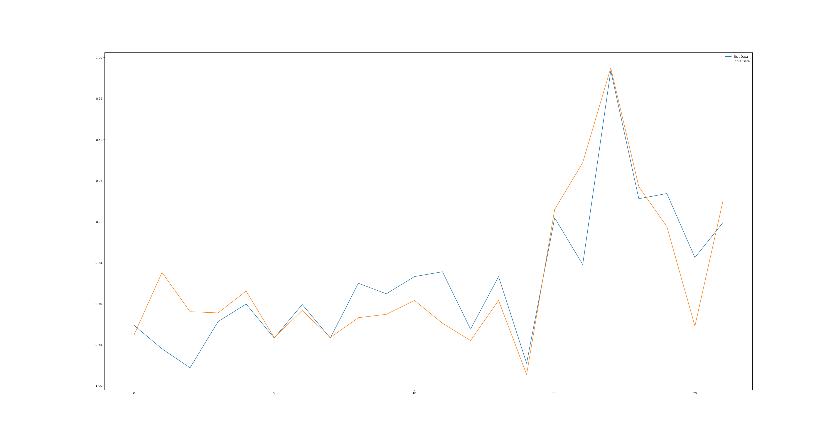
# Part D: Post-implementation Report

## A Business (or Organization) Vision

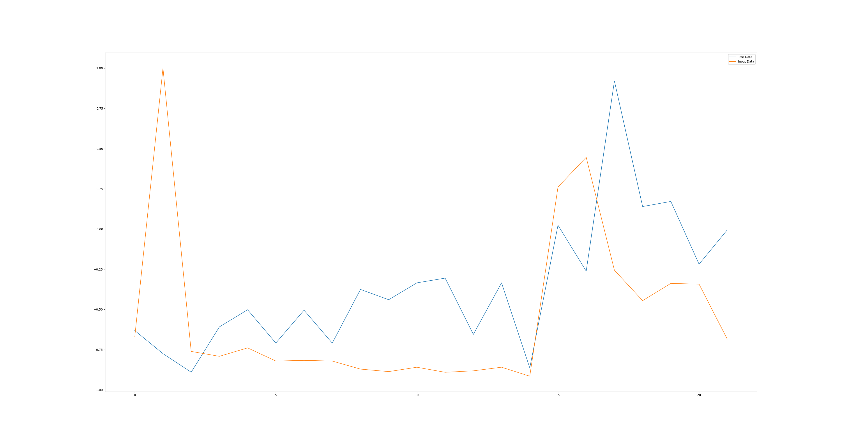
Detecting Parkinson’s Disease, or PD, as early as possible is crucial for patients. When PD is in it’s early stages, it can be very difficult to detect by humans. With this application, we are able to detect PD early by using patient vocal recording data. Doctors can use this application, along with a vocal recording system, to accuratly detect and diagnose PD in their patients. Doctors just have to type in their patients data and the application will be able to provide it’s prediction on whether or not it detects PD and give an accuracy rating on how sure the application is. For example, this is the graph that the input data is being compared to:



Here is a positive test result:



Here is a negative test result:



## Datasets

The dataset used was provided by Oxford University and University of California, Irvine (UCI). You can find this dataset at the following website:

<https://archive.ics.uci.edu/ml/datasets/parkinsons>

This data set contains 197 patients biomedical vocal data. There are a total of 31 people that participated in this dataset, and 23 are confirmed with PD. Each column in the table represents a particular voice measure. The dataset is provided in a CSV file format. The dataset in this format made it easily accessible to my algorithm. The following is an example of the raw data:

phon\_R01\_S01\_1,119.99200,157.30200,74.99700,0.00784,0.00007,0.00370,0.00554,0.01109,0.04374,0.42600,0.02182,0.03130,0.02971,0.06545,0.02211,21.03300,1,0.414783,0.815285,-4.813031,0.266482,2.301442,0.284654

Some data, like the patient name and status, are not used in the algorithm, so the processed data looks more like this:

MDVP:Fo(Hz): 116.01400

MDVP:Fhi(Hz): 141.78100

MDVP:Flo(Hz): 110.65500

MDVP:Jitter(%): 0.01284

MDVP:Jitter(Abs): 0.00011

MDVP:RAP: 0.00655

MDVP:PPQ: 0.00908

Jitter:DDP: 0.01966

MDVP:Shimmer: 0.06425

MDVP:Shimmer(dB): 0.58400

Shimmer:APQ3: 0.03490

Shimmer:APQ5: 0.04825

MDVP:APQ: 0.04465

Shimmer:DDA: 0.10470

NHR: 0.01767

HNR: 19.64900

RPDE: 0.417356

DFA: 0.823484

Spread1: -3.747787

Spread2: 0.234513

D2: 2.332180

PPE: 0.410335

## Data Product Code

To better understand the data and how it is being used, we have created a GUI environment to input the data and to show the results of the data. This GUI environment is provided by the PySimpleGUI library and the graph is provided by the MatPlotLib library. The dataset provided by UCI is loaded into the program using a CSV reader. This data is shown in the console of the IDE as the program is running to give a better understanding of the dataset. The data is then split between the features and labels. The features are then scaled from -1 to 1. Next the dataset is split. 80% of the dataset is used to train the model and the remaining 20% is used to test the model. The model is trained using the XGBClassifier model and using the trained model, we can show the accuracy of the predictions with this model is 94%.

When data is inputted into this program, the input data is used in the model to predict if this patient is showing signs of PD or not. To better understand how the data is being compared, a graph is provided. The blue line in the graph represents a positive PD test result. The orange line in the graph represents the input data. This comparison, along with others within the program, are used to accurately predict the presence of PD in a patient.

## Objective (or Hypothesis) Verification

The project’s hypothesis is that it would provide a greater than 80% accuracy when detecting Parkinson’s Disease using vocal data. This project successfully meets this hypothesis. This project provides a 94% accuracy while detecting Parkinson’s Disease.

## Effective Visualization and Reporting

The visual elements of our program are used to compare positive test results to the input data. This comparison is used to help show the accuracy of the data and to better understand how the data is being read. I have also included the full dataset in the consol of the program as a way of showing the full exploration of the data. The input data consists of the following information:

name - ASCII subject name and recording number

MDVP:Fo(Hz) - Average vocal fundamental frequency

MDVP:Fhi(Hz) - Maximum vocal fundamental frequency

MDVP:Flo(Hz) - Minimum vocal fundamental frequency

MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several measures of variation in fundamental frequency

MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of variation in amplitude

NHR,HNR - Two measures of ratio of noise to tonal components in the voice

status - Health status of the subject (one) - Parkinson's, (zero) - healthy

RPDE,D2 - Two nonlinear dynamical complexity measures

DFA - Signal fractal scaling exponent

spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

The input data is being compared to the following graph:

Chart, line chart

Description automatically generated

Here is what a positive test result looks like compared to the original graph:

Chart, line chart

Description automatically generated

Here is what a negative test result looks like compared to the original graph:

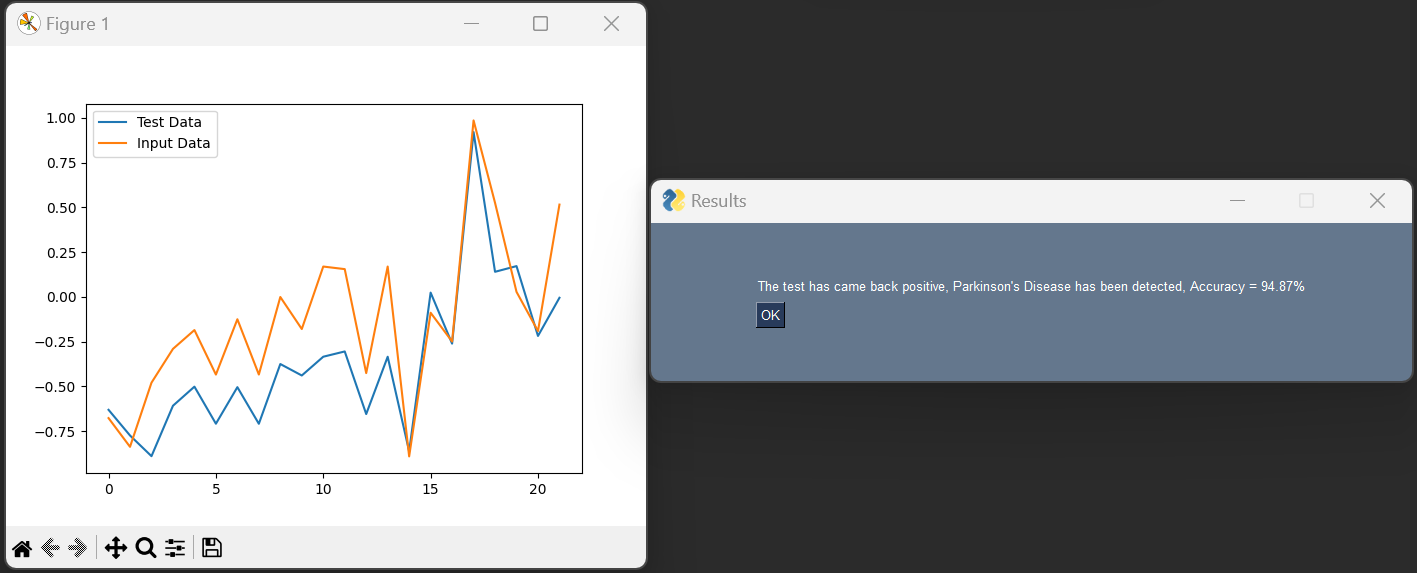
Chart, line chart

Description automatically generated

The positive test result more closely follows the patterns of the original graph and that is used to determine that this is a positive test. The negative test doesn’t follow the original graph as well, so the program will determine that this is a negative test.

## Accuracy Analysis

For this accuracy test, I used Classification Accuracy testing. To find the accuracy, I used the built in accuracy score method for SKLearn. I used the variables of the total results being tested and the amount of correct predictions to find the accuracy. This accuracy is shown in the results section of the program. The following is an example of what it looks like with a positive test result:



## Application Testing

To train this program, I split the dataset into 20% for testing and the other 80% for training. This application was tested by using data from the UCI Parkinson’s Dataset. I would choose patients that were confirmed positive for PD and patients confirmed negative for PD and would plug their information into the program. The application was able to accuratly detect PD in the positive patients and didn’t detect it in the confirmed negative patients. This application started at a 94% accuracy, so I saw no modification necessary for this application.

## Application Files

For this program, you will need the Capstone Project file and the following libraries for Python:

1. Pandas
2. NumPy
3. PySimpleGUI
4. SKLearn
5. MatPlotLib
6. XGBoost

Here is the file organization:

1. Capstone Project
   1. .idea – PyCharm Files
      1. InspectionProfiles
         1. Profiles\_settings.xml
         2. Project\_Default.xml
      2. .gitignore
      3. .name
      4. Capstone.iml
      5. Modules.xml
      6. Workspace.xml
   2. Venv – PyCharm Files
      1. Lib
      2. Scripts
      3. .gitignore
      4. Pyvenv.cfg
   3. Main.py – Main Program
   4. Parkinsons.data - Dataset
   5. Test Data.docx – Testing data for evaluators

## User Guide

For evaluators and users trying to start this application, please follow the instructions below:

1. Unzip the associated files and place them into a project folder
2. Ensure that you have Python 3.10 installed on the desktop that you are using for the program.
3. Load the project folder into an IDE that can support Python 3.10 (We recommend using JetBrains PyCharm Community Edition 2022.3.3)
4. Download and install the pandas, numpy, PySimpleGUI, sklearn, matplotlib, and xgboost libraries.
   1. This can be done by opening your local command prompt and entering the following without the quotation marks: “pip install pandas numpy pysimplegui sklearn matplotlib xgboost”
5. Navigate to the python file called “main.py” and start the application by right clicking the file and selecting “Run”.
6. Once the application starts, a window will pop up asking for the vocal measurements for your patient. Enter all of the information from the vocal measurements into the program. I will include test data (Document named Test Data) as examples to plug in and test the program.
7. After the information is entered, press the “Submit” button. A window will pop up if the data has been read successfully.
8. After the window pops up, a graph will show you how closely the patient data matches to confirmed positive Parkinson’s Disease data. After that graph is shown, a window will pop up with the results and the accuracy of the results.

## Summation of Learning Experience

My previous experiences through WGU and working with BlumeTech have prepared me for this project. I used my experiences making project proposals and working in Python and artificial intelligence. I also had experience in the medical field that helped me with understanding the vocal data. This project experience makes me want to continue learning more about medical data and other ways that I can use my coding expertise to help people. I want to continue learning and provide the best experience for all of my clients.